**Project 2:- Create a Linear Regression Model for DVD sales Data set?**

library(readxl)

Sales\_dataset <- read\_excel("C:/Users/Anjana/Desktop/Sales\_dataset.xlsx")

View(Sales\_dataset)

DVDsales<- lm(sales~advertise+plays+attractiveness,data=Sales\_dataset)

summary(DVDsales)

Call:

lm(formula = sales ~ advertise + plays + attractiveness, data = Sales\_dataset)

Residuals:

Min 1Q Median 3Q Max

-122.728 -28.760 1.476 29.422 142.960

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -28.140377 17.373604 -1.62 0.107

advertise 0.084642 0.006908 12.25 < 2e-16 \*\*\*

plays 3.385493 0.277723 12.19 < 2e-16 \*\*\*

attractiveness 11.333342 2.437340 4.65 6.1e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.1 on 196 degrees of freedom

Multiple R-squared: 0.6645, Adjusted R-squared: 0.6593

F-statistic: 129.4 on 3 and 196 DF, p-value: < 2.2e-16

**Output Interpretation-** From this output, we have determined that the intercept is -28.140377 and the coefficient of advertise is 0.084642, for plays it is 3.385493 and for attractiveness is 11.333342.

model<- -28.140377 + (0.084642\* advertise)+(3.385493\*plays)+(11.333342\*attractiveness)

This equation tells us that the predicted number of DVD sales will increase by 0.084642 for every one percent increase in the advertise, will increase by 3.385493 for every one percent increase in the plays and will increase by 11.333342 for every one percent increase in the attractiveness.

The Adjusted R-squared value is 0.6593 which means that the model can explain about 66% of the variance of the Sales variable.

**Obtaining the confidence bands:-**

predict(DVDsales, interval="confidence")

fit lwr upr

1 231.63734 211.59237 251.68232

2 229.41738 221.05928 237.77547

3 292.04091 278.60949 305.47234

4 263.48556 253.01375 273.95738

5 226.11608 211.12239 241.10977

6 141.00822 129.93634 152.08011

7 90.83821 62.81115 118.86527

8 193.82321 180.40252 207.24389

9 165.80022 158.08818 173.51226

10 201.34836 189.89791 212.79881

11 305.18161 288.72130 321.64192

12 113.99310 90.48527 137.50093

13 165.03964 154.49909 175.58019

14 176.80098 164.56742 189.03455

15 166.87849 159.40874 174.34823

16 135.62753 124.24014 147.01493

17 259.02452 243.46792 274.58112

18 201.04900 189.66473 212.43328

19 266.22085 253.24155 279.20015

20 291.11219 277.72086 304.50352

**Fitted Values and Residuals:-**

sales\_fitted<-data.frame(Sales\_dataset, fitted.value=fitted(DVDsales),residual=resid(DVDsales))

> sales\_fitted

advertise sales plays attractiveness fitted.value residual

1 10.256 330 43 10 231.63734 98.3626573

2 985.685 120 28 7 229.41738 -109.4173751

3 1445.563 360 35 7 292.04091 67.9590854

4 1188.193 270 33 7 263.48556 6.5144353

5 574.513 220 44 5 226.11608 -6.1160825

6 568.954 170 19 5 141.00822 28.9917772

7 471.814 70 20 1 90.83821 -20.8382065

8 537.352 210 22 9 193.82321 16.1767937

9 514.068 200 21 7 165.80022 34.1997792

10 174.093 300 40 7 201.34836 98.6516378

11 1720.806 290 32 7 305.18161 -15.1816078

12 611.479 70 20 2 113.99310 -43.9931010

13 251.192 150 24 8 165.03964 -15.0396392

14 97.972 190 38 6 176.80098 13.1990150

15 406.814 240 24 7 166.87849 73.1215138

16 265.398 100 25 5 135.62753 -35.6275350

17 1323.287 250 35 5 259.02452 -9.0245217

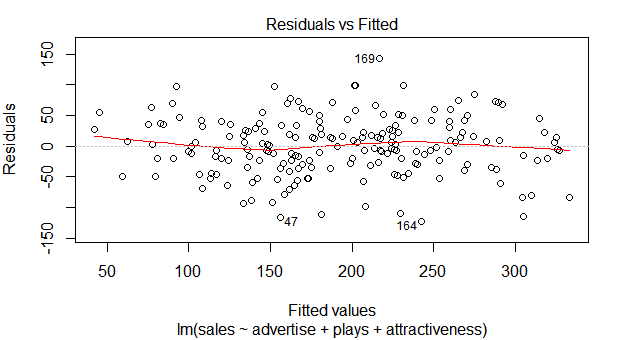
18 196.650 210 36 8 201.04900 8.9509955

19 1326.598 280 27 8 266.22085 13.7791501

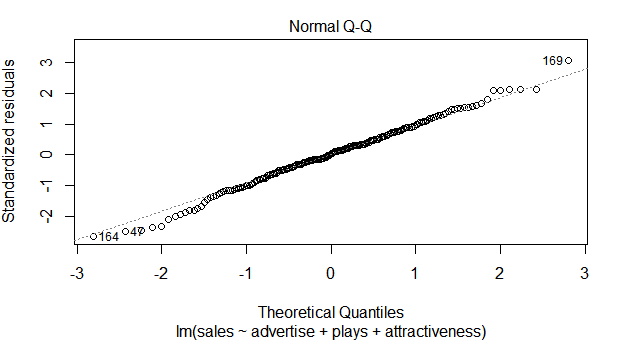
20 1380.689 230 33 8 291.11219 -61.1121913

**Diagnostic Plots:-**

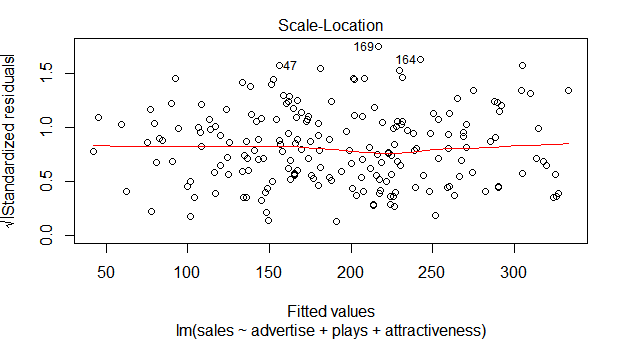
plot(DVDsales)



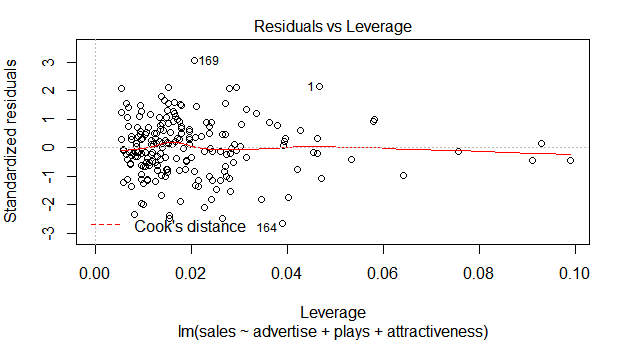
This plot shows if residuals have non-linear patterns. Here we can see equally distributed residuals around the horizontal line without distinct pattern. It means it does not have non-linear relationship.



This plot shows if residuals are normally distributed. Deviations from a straight line could mean that the errors which do not follow a normal distribution. Here you can a few outliers.



This plot shows if residuals are spread equally along the ranges of predictors. Its good to have a horizontal line.



This plot helps to find out influential cases which are determined to a regression model. That means, the results wouldn’t be much different if we either include or exclude them from analysis.